Density Estimation by Minimum Encoding Mixtures of Histograms

Petri T. Kontkanen, Petri J. Myllymäki*, Tomi V. Silander, Henry R. Tirri
University of Helsinki, Finland.

cosco@cs.Helsinki.FI

http://www.cs.Helsinki.FI/research/cosco

Estimating the density function of a continuous random variable is a central problem in many predictive modeling tasks where the goal is to construct from sample data a domain model that can be used for predicting future events. Nonparametric *kernel estimators* provide one possible solution to this problem. The term "nonparametric" means in this context that the density function to be estimated is not assumed to have any pre-specified functional form. However, this does not mean that nonparametric density estimators do not contain any parameters; on the contrary, nonparametric models quite often contain a huge number of parameters when compared to parametric models. What is more, although clever techniques for tuning these parameters exist (Scott, 1992; Wand & Jones, 1995), finding reasonable values for the parameters can turn out to be very difficult in practice.

An alternative approach for density estimation is offered by the Bayesian modeling approach (Bernardo & Smith, 1994; Gelman, Carlin, Stern, & Rubin, 1995). In this framework, a certain family of possible model structures is assumed, and the choice between different model structures is based on *Bayes factors*, obtained by computing the posterior probability of each model structure given the sample data. However, the posterior probability of a model structure involves integration over all the possible model parameter instantiations, which in many cases can be computationally infeasible. What is more, even if the posterior could be computed efficiently, this approach requires the user to provide certain prior distributions for the model parameters, as well as for the model structures. In non-informative cases (in situations where no prior knowledge of the problem domain is available), determining these prior distributions in a theoretically reasonable manner is a highly controversial issue, as discussed in, e.g., (Berger & Bernardo, 1992; Ghosh & Mukerjee, 1992; Hartigan, 1996; Hjort, 1996; Andreev & Arjas, 1996; Cowell, 1992).

Information theory offers a third, alternative framework for density estimation. One of the main advantages of this approach is that it is totally parameter-free: the only thing that matters is the sample data available, and all modeling problems are solved by using information-theoretic formulations of optimality. The *minimum description length (MDL)* principle is perhaps the best known attempt to formalize this framework. This principle aims at finding an efficient coding for all reasonable data sequences. Intuitively speaking, the approach is based on the idea that the more we are able to compress a given set of data, the more we have learned about the domain the data was collected from. However, although such informal descriptions of the MDL principle are intuitively quite appealing, they should not be taken too seriously — formally, the principle can be defined via an elaborate theoretical framework developed in (Rissanen, 1987, 1989; Clarke & Barron, 1990; Rissanen, 1996; Takeuchi & Barron, 1998). A recent, comprehensive tutorial to the MDL theory and its application to predictive modeling can be found in (Grünwald, 1998).

Unfortunately, many reported applications of MDL found in the literature use ad hoc methods based on straightforward applications of the intuitive principle, rather than the actual theoretical framework and the corresponding results. Another problem in applying MDL in practice lies in the fact that the theoretical framework has gone through several evolutionary steps during the last two decades, which has caused some confusion in this research area. For this reason, many researchers tend to use more widely known, but "outdated" versions of MDL, such as the two-part MDL estimator (Rissanen, 1978), also known as the Schwarz or BIC criterion (Schwarz, 1978)), although much more efficient techniques have been developed recently.

In this paper we concentrate on an advanced MDL approach based on the *normalized maximum likelihood estimator* (NML) discussed in (Rissanen, 1996). In particular, we study the novel density model suggested by Barron et al. (Barron, Rissanen, & Yu, 1998), based on a finite mixture of histogram models where the mixture coefficients are determined by the normalized maximum likelihood measure. Theoretically, the NML mixture density produces a very accurate density estimator regardless of the properties of the data generating process. However, not counting the two simple experiments reported in (Barron et al., 1998), to our knowledge this density estimator has not been empirically validated. We present results of extensive empirical tests where the accuracy of the NML density is compared to the accuracy of alternative density estimators, by using several different real-world data sets. The preliminary results suggest that the NML density estimator provides a robust, parameter-free approach to density estimation, which yields good results over a wide range of different problem domains.

References

Andreev, A., & Arjas, E. (1996). A note on histogram approximation in Bayesian density estimation. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics 5* (pp. 487–490). Oxford University Press.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description principle in coding and modeling. *IEEE Transactions on Information Theory, 44*(6), 2743–2760.

Berger, J., & Bernardo, J. (1992). On the development of reference priors. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics 4* (pp. 35–60). Oxford University Press.

Bernardo, J., Berger, J., Dawid, A., & Smith, A. (Eds.). (1992). *Bayesian statistics 4.* Oxford University Press.

Bernardo, J., Berger, J., Dawid, A., & Smith, A. (Eds.). (1996). *Bayesian statistics 5.* Oxford University Press.

Bernardo, J., & Smith, A. (1994). *Bayesian theory.* John Wiley.

Clarke, B., & Barron, A. (1990). Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory, 36*(3), 453–471.

Cowell, R. (1992). On compatible priors for Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(9), 901–911.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1995). *Bayesian data analysis.* Chapman & Hall.

Ghosh, J., & Mukerjee, R. (1992). Non-informative priors. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics 4* (pp. 195–210). Oxford University Press.

Grünwald, P. (1998). *The minimum description length principle and reasoning under uncertainty.* Unpublished doctoral dissertation, CWI, ILLC Dissertation Series 1998-03.

Hartigan, J. (1996). Bayesian histograms. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics 5* (pp. 211–222). Oxford University Press.

Hjort, N. (1996). Bayesian approaches to non- and semiparametric density estimation. In J. Bernardo, J. Berger, A. Dawid, & A. Smith (Eds.), *Bayesian statistics 5* (pp. 223–253). Oxford University Press.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*, 445-471.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society*, *49*(3), 223–239 and 252–265.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* New Jersey: World Scientific Publishing Company.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *42*(1), 40–47.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Scott, D. (1992). *Multivariate density estimation. theory, practice, and visualization.* New York: John Wiley & Sons.

Takeuchi, J., & Barron, A. (1998). Asymptotically minimax regret by Bayes mixtures. In *1998 IEEE international symposium on information theory.* Cambridge, MA.

Wand, M., & Jones, M. (1995). *Kernel smoothing.* Chapman & Hall.

Complex Systems Computation Group, Department of Computer Science,
P.O.Box 26, FIN-00014 University of Helsinki, Finland.